ENJEUX ET DÉFIS DES DONNÉES

GÉOSPATIALES POUR L'ONCOLOGIE

GROUPE GEOCANCER

Cancéropôle Île-de-France

RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DU CANCER

Etude de biais sur les adresses de patients hospitalisés à l'HEGP et l'Institut Curie

Joséphine Bocquet – Ingénieure Géomaticienne

Enjeux et défis des données géospatiales pour l'oncologie

Ensemble des adresses et cartes présentés sont des **données simulées**

pour préserver la confidentialité des patients

Enjeux et défis des données géospatiales pour l'oncologie

Objectif : **Géolocalisation** des adresses patients pour enrichir les données de santé avec des éléments annexes

Entrepôt de données de santé

ID Patient	Adresse patient	Autres données …
1	20 rue Leblanc 75015 Paris	…
2	10 rue des Terres au Curé 75013 Paris	…

Exemple de table extraites de l'entrepôt

Enjeux et défis des données géospatiales pour l'oncologie

Objectif : **Géolocalisation** des adresses patients pour enrichir les données de santé avec des éléments annexes

ID Patient	Adresse patient
1	20 rue Leblanc 75015 Paris
2	10 rue des Terres au Curé 75013 Paris

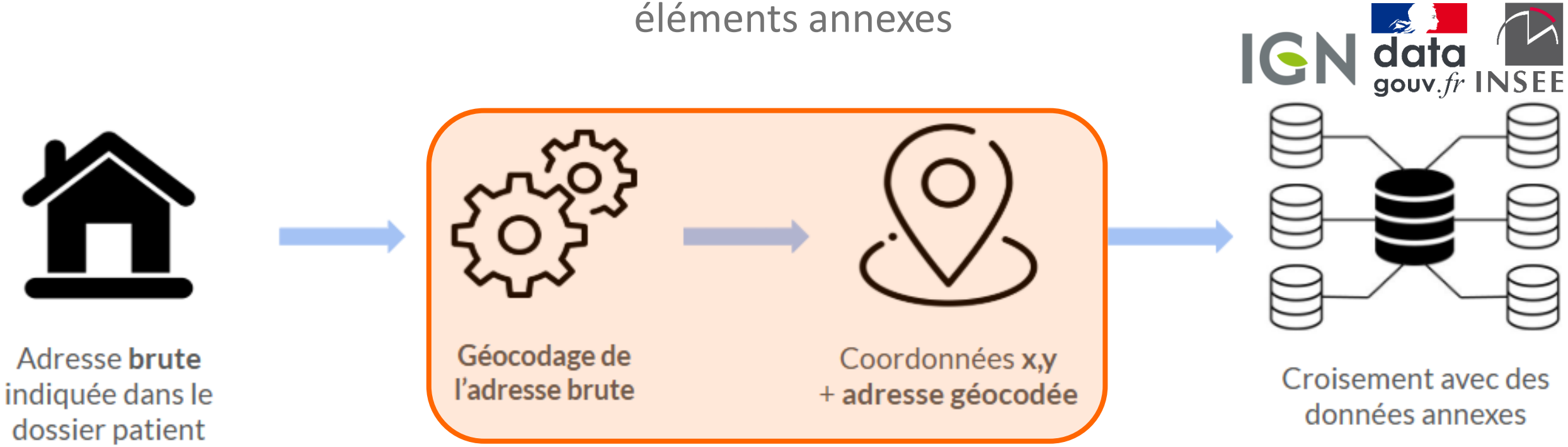
x Environ 2 millions de patients

Adresse **brute** indiquée dans le dossier patient

Géocodage de l'adresse brute

Enjeux et défis des données géospatiales pour l'oncologie

Objectif : **Géolocalisation** des adresses patients pour enrichir les données avec des éléments annexes



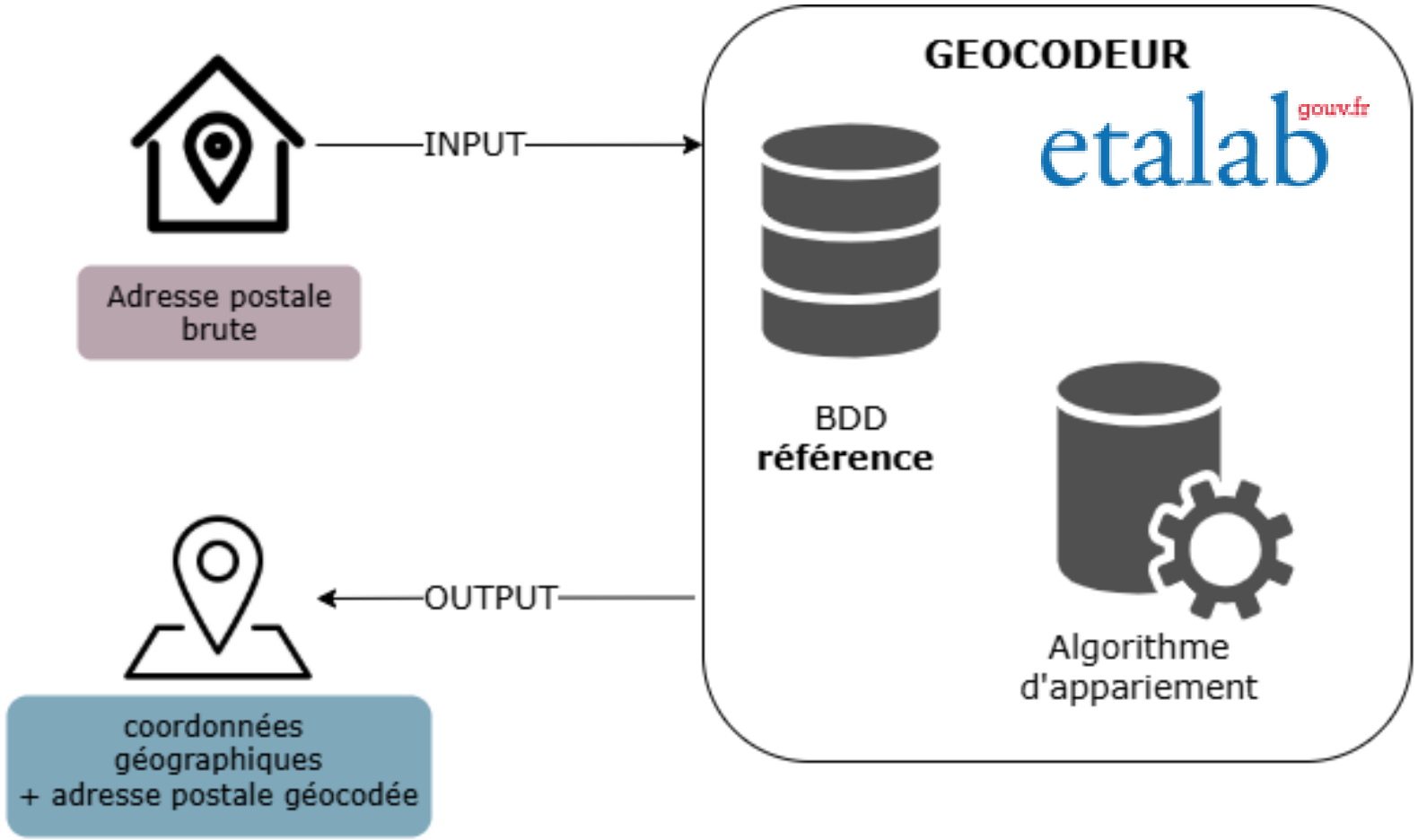
Problématique : **Qualité** des adresses fournies par les patients

Des données géographiques à l'hôpital

Géocodage

Adresse initiale

10 rue des Terres au Curé
75013 Paris



Schématisation du fonctionnement d'un géocodeur

Des données géographiques à l'hôpital

Géocodage

Adresse initiale

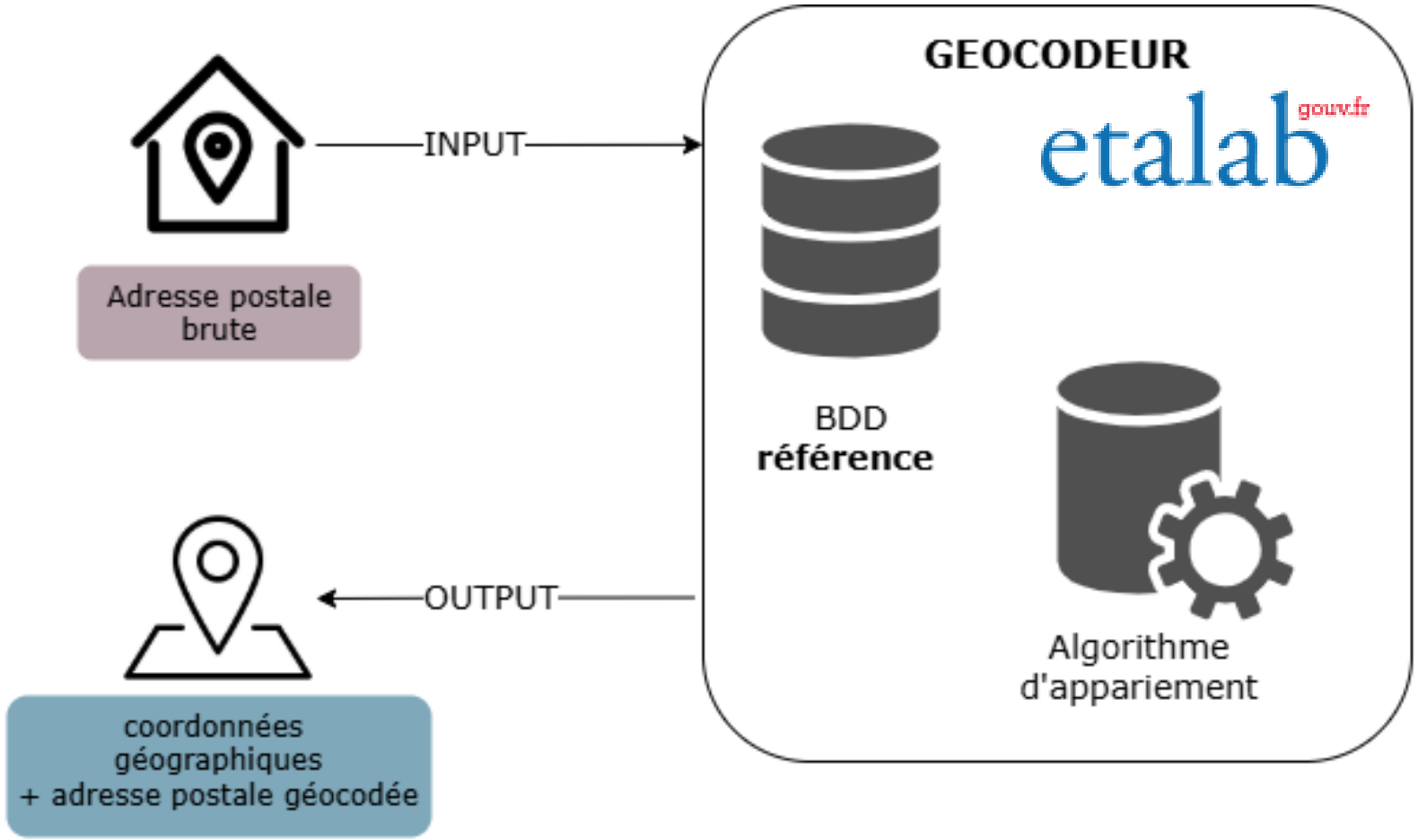
10 rue des Terres au Curé
75013 Paris

Adresse géocodée

10 rue des Terres au Curé
75013 Paris

Coordonnées géographiques

y : 48.82413102561786
x : 2.372370381468407



Schématisation du fonctionnement d'un géocodeur

Des données géographiques à l'hôpital

Problématique : Qualité des adresses fournies par les patients

Adresses postales correctes :

- Eléments additionnels (Esc, Appt, Bâtiment, etc.)
- Adresse administrative camouflée

Adresses postales incorrectes :

- Adresses relatives (« Chez Mr », etc.)
- Libellé administratif (Centre hospitalier, pénitentiaires, etc.)

Des données géographiques à l'hôpital

Exemples d'adresses retrouvées :

Pour l'adresse : **10 rue des Terres au Curé 75013 Paris**

On peut trouver des erreurs du type :

Eléments additionnels **Appartement T Bat C 10 rue des Terres au Curé 75013 Paris**

Des données géographiques à l'hôpital

Exemples d'adresses retrouvées :

Pour l'adresse : **10 rue des Terres au Curé 75013 Paris**

On peut trouver des erreurs du type :

Eléments additionnels	**Appartement T Bat C 10 rue des Terres au Curé 75013 Paris**
Libellé administratifs	**Maison des Associations de Solidarité 75013 Paris**

Des données géographiques à l'hôpital

Exemples d'adresses retrouvées :

Pour l'adresse : **10 rue des Terres au Curé 75013 Paris**

On peut trouver des erreurs du type :

Eléments additionnels	**Appartement T Bat C 10 rue des Terres au Curé 75013 Paris**
Libellé administratifs	**Maison des Associations de Solidarité 75013 Paris**
Adr. administrative camouflée	**10 rue des Terres au Curé 75013 Paris**

Des données géographiques à l'hôpital

Problématique : *Qualité de l'écriture des adresses*

Adresses postales correctes :

Eléments additionnels

Adr. administrative camouflée

Adresses postales incorrectes :

Libellé administratifs

Adresses relatives

On retrouve donc **2 types de biais** : des **biais textuels** et des **biais géographiques**

Des données géographiques à l'hôpital

Problématique : *Qualité de l'écriture des adresses*

Adresses postales correctes :

Eléments additionnels

Adr. administrative camouflée

Adresses postales incorrectes :

Libellé administratifs

Adresses relatives

Adresses géocodables

On retrouve donc **2 types de biais** : des **biais textuels** et des **biais géographiques**

Identification des biais textuels

Exemple

1 . RAW DATA

raw address : Hopital Europeen George Pompidou 20 rue Leblanc

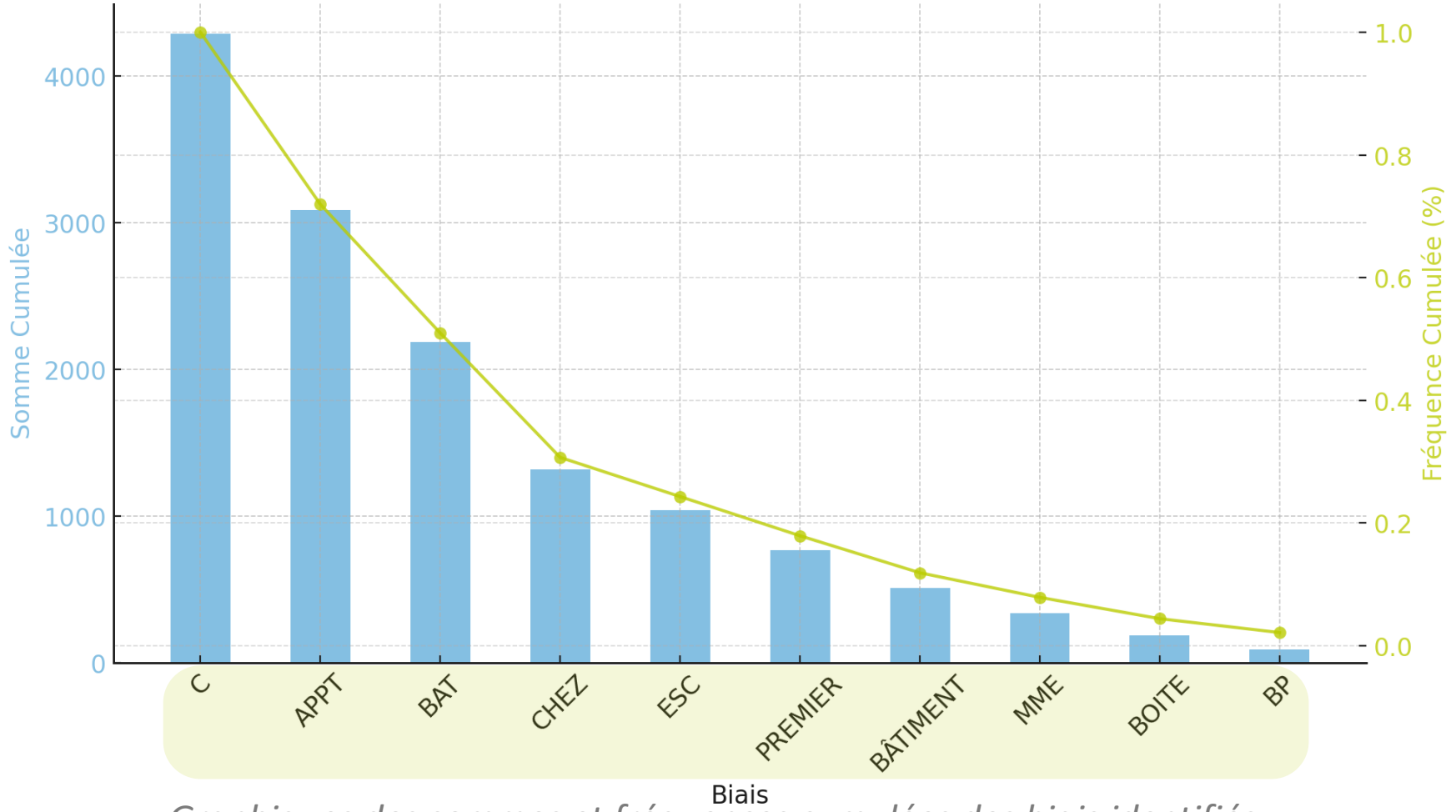
geocoded address : 20 rue Leblanc

2. SEQUENCE ALIGNMENT

```
Hopital HEGP 20 rue Leblanc
-------------20 rue Leblanc
```

Identification des biais textuels

Résultat : **10 éléments additionnels les plus fréquents**



Graphiques des sommes et fréquences cumulées des biais identifiés

Les 10 biais identifiés sont présents dans :

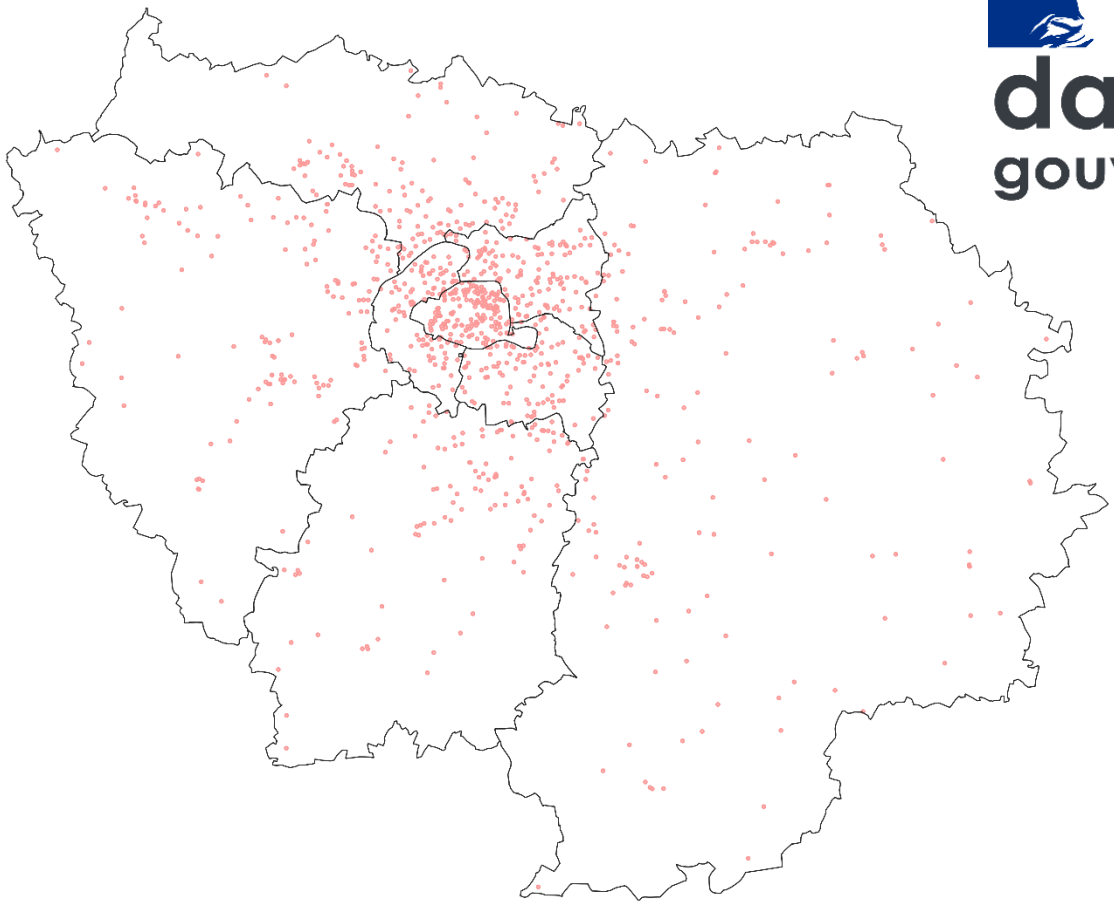
4397 adresses

soit **0.238** %

Simulation de biais sur des données de références

Adresse et **coordonnées**
des écoles

Adresse école	Coordonnées x,y
17 rue Auguste Compte 75006 Paris	x : 2.3348 y : 48.8443



Répartition de mille écoles du premier et second degrés en Île-de-France

Simulation de biais sur des données de références

Intégration des biais

Adresse et **coordonnées** de référence

Adresse et **coordonnées** de ref
+ **Adresse** de **référence biaisée**

Adresse école réf	Coordonnées ref x,y
17 rue Auguste Compte 75006 Paris	x : 2.3348 y : 48.8443

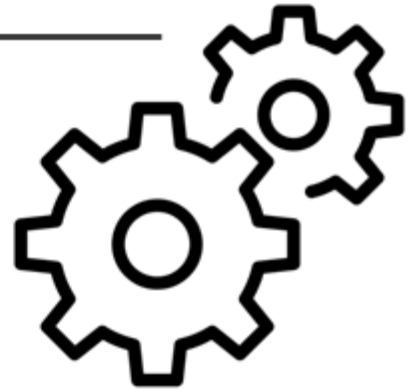
Adresse école réf	Coordonnées ref x,y	Adresse école réf biaisée
17 rue Auguste Compte 75006 Paris	x : 2.3348 y : 48.8443	APPT 17 rue Auguste Compte 75006 Paris

Simulation de biais sur des données de références

Intégration des **biais**

Adresse et **coordonnées** de référence

Adresse et **coordonnées** de ref
+ **Adresse** de **référence biaisée**



Géocodage de l'adresse de référence biaisée

Adresse et **coord** de référence
+ **Adresse** et **coord** de **réf biaisée**

Adr. école réf	Coordonnées x,y ref	Adresse école réf biaisée	Coordonnées x,y ref biaisée
17 rue Auguste Compte 75006 Paris	x : 2.3348 y : 48.8443	**APPT** 17 rue Auguste Compte 75006 Paris	x : 2.4444 y : 48.8888

Schéma d'intégration des biais **17**

Simulation de biais sur des données de références

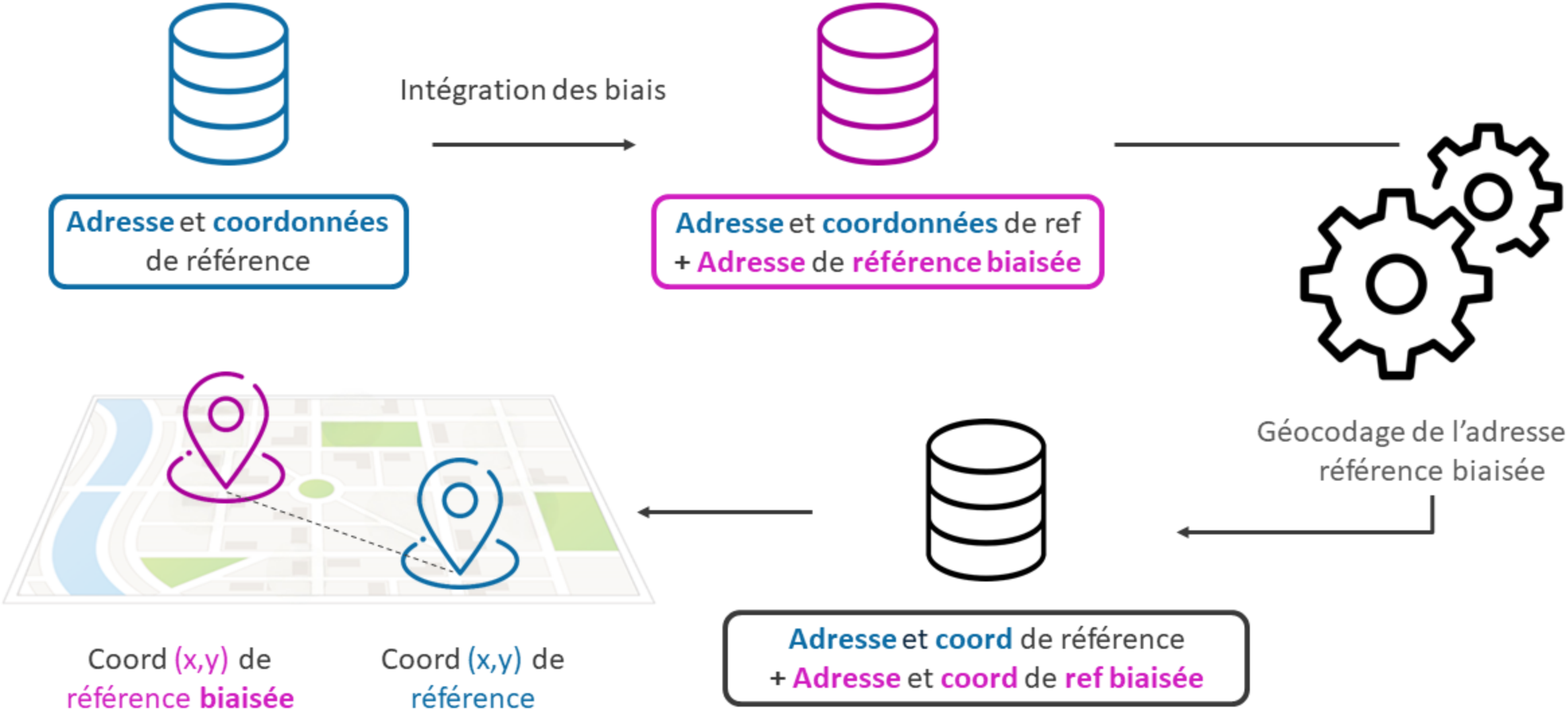
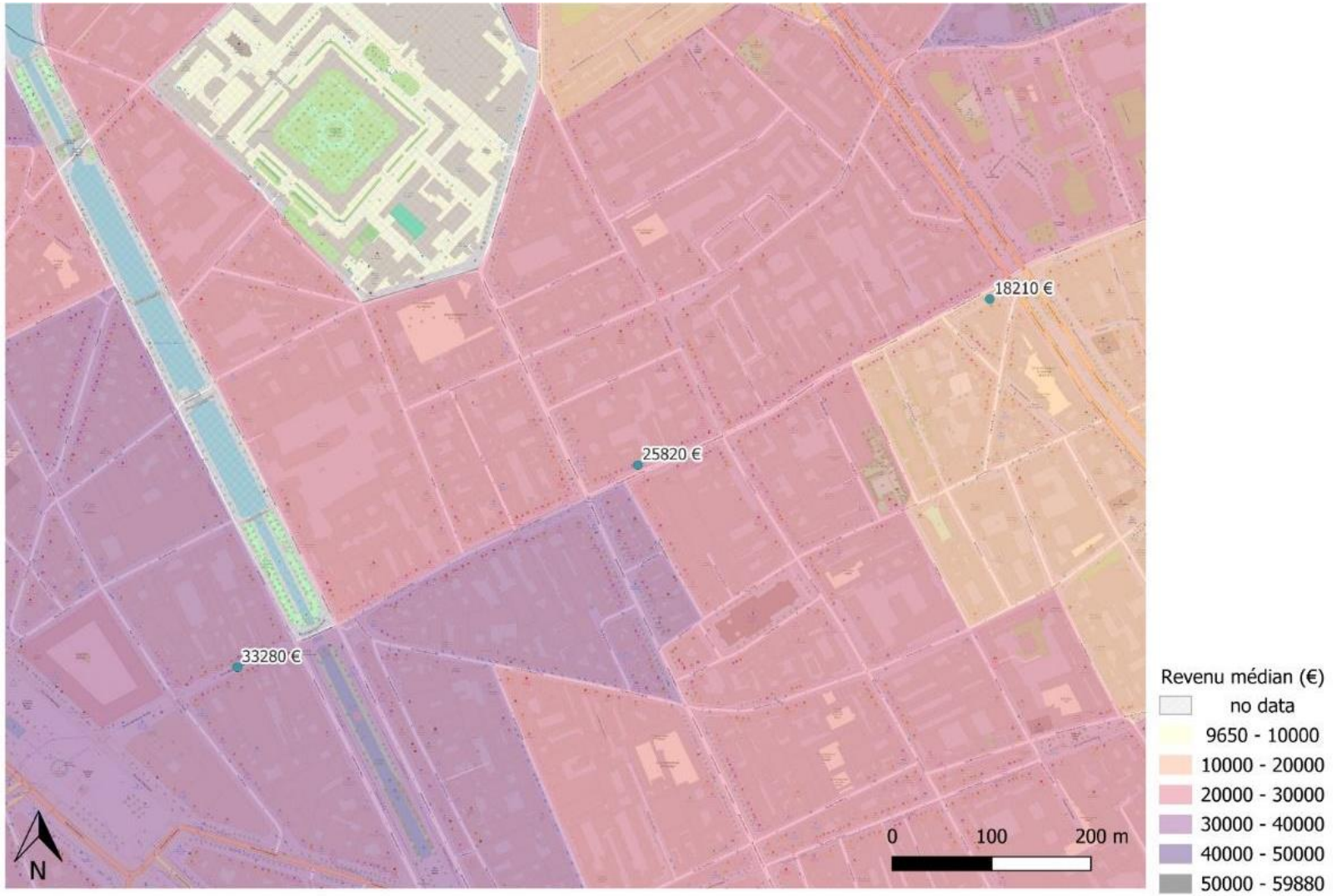


Schéma d'intégration des biais et mesure de l'impact

Simulation de biais sur des données de références

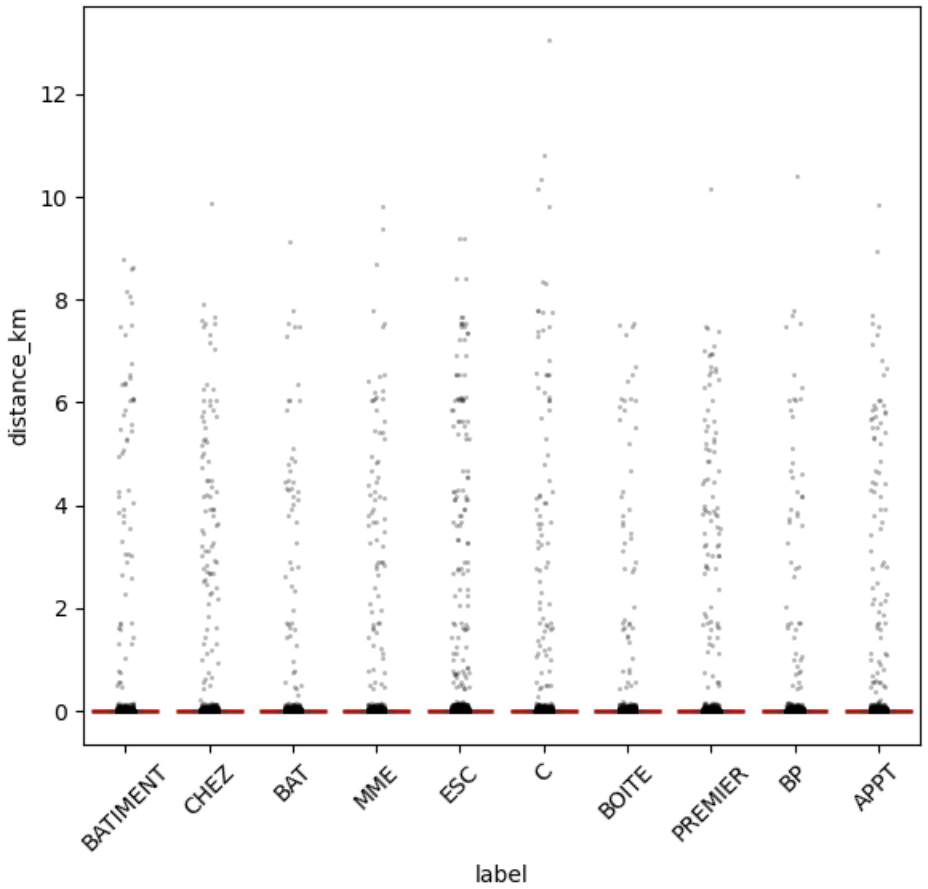
Mesure de l'impact sur des données socioéconomiques

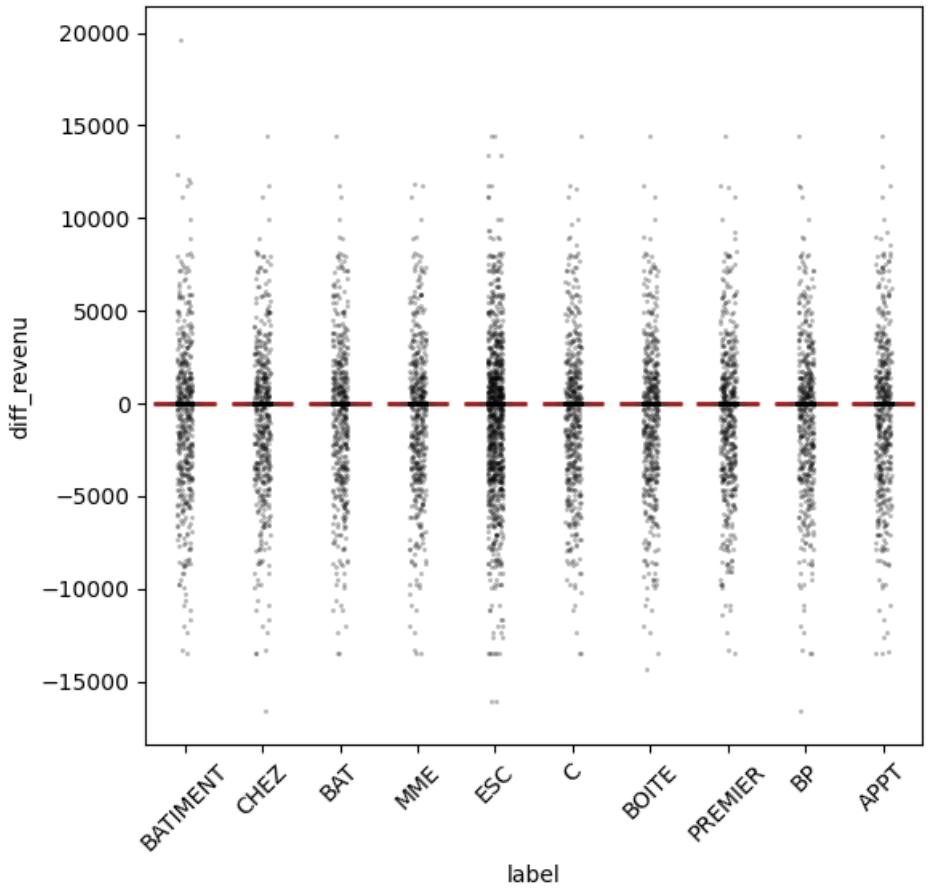


Distribution du revenu médian par unité de consommation le long de la rue du Faubourg Saint Antoine

Source : insee.com Dispositif Fichier localisé social et fiscal (Filosofi)

Simulation de biais sur des données de références



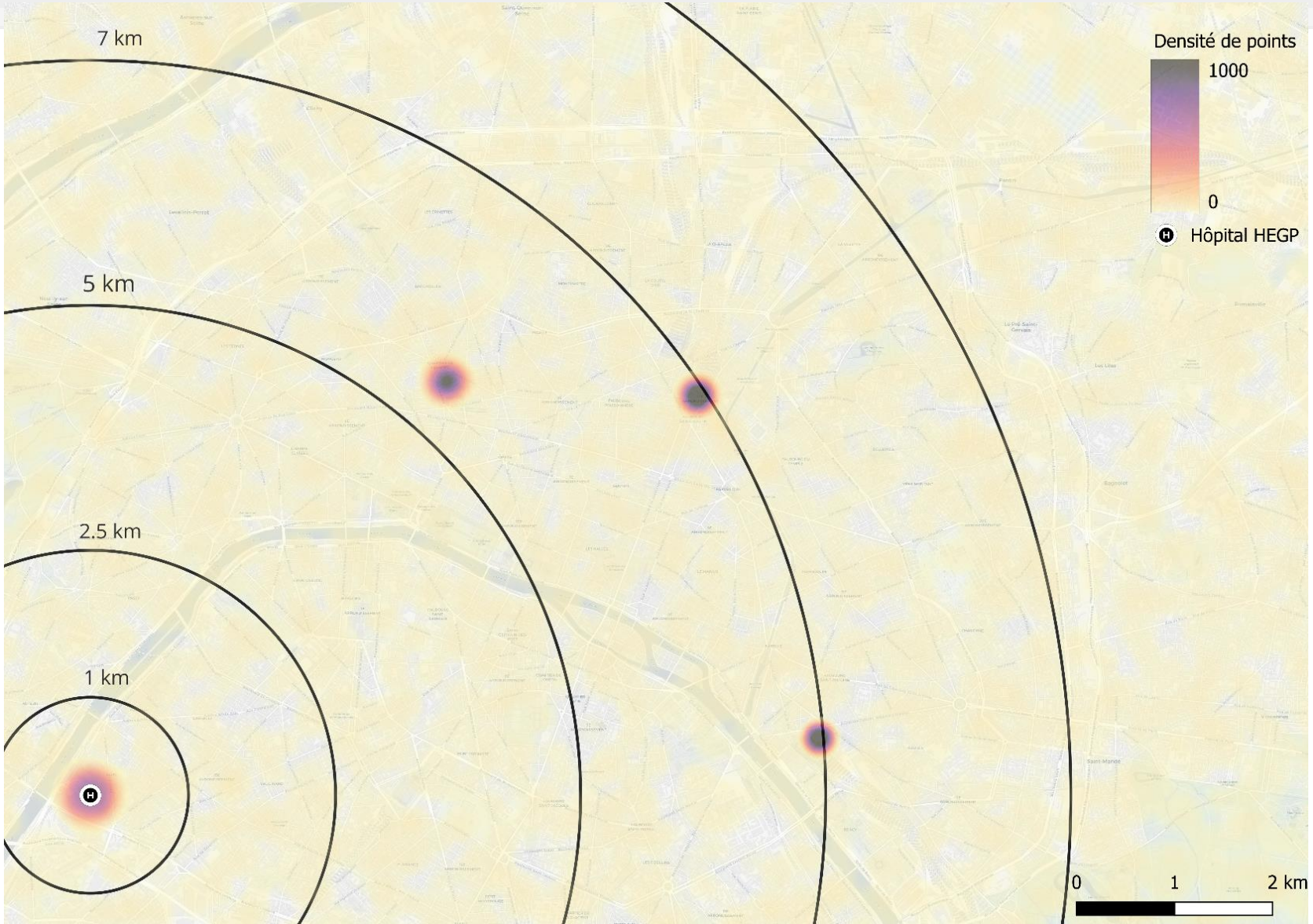


Figures de la distance (à gauche) et différence de revenu (à droite) entre les coordonnées de références et les coordonnées de référence biaisées

Conclusion

Géocodeur très **robuste** aux éléments additionnels

Identification des biais géographiques



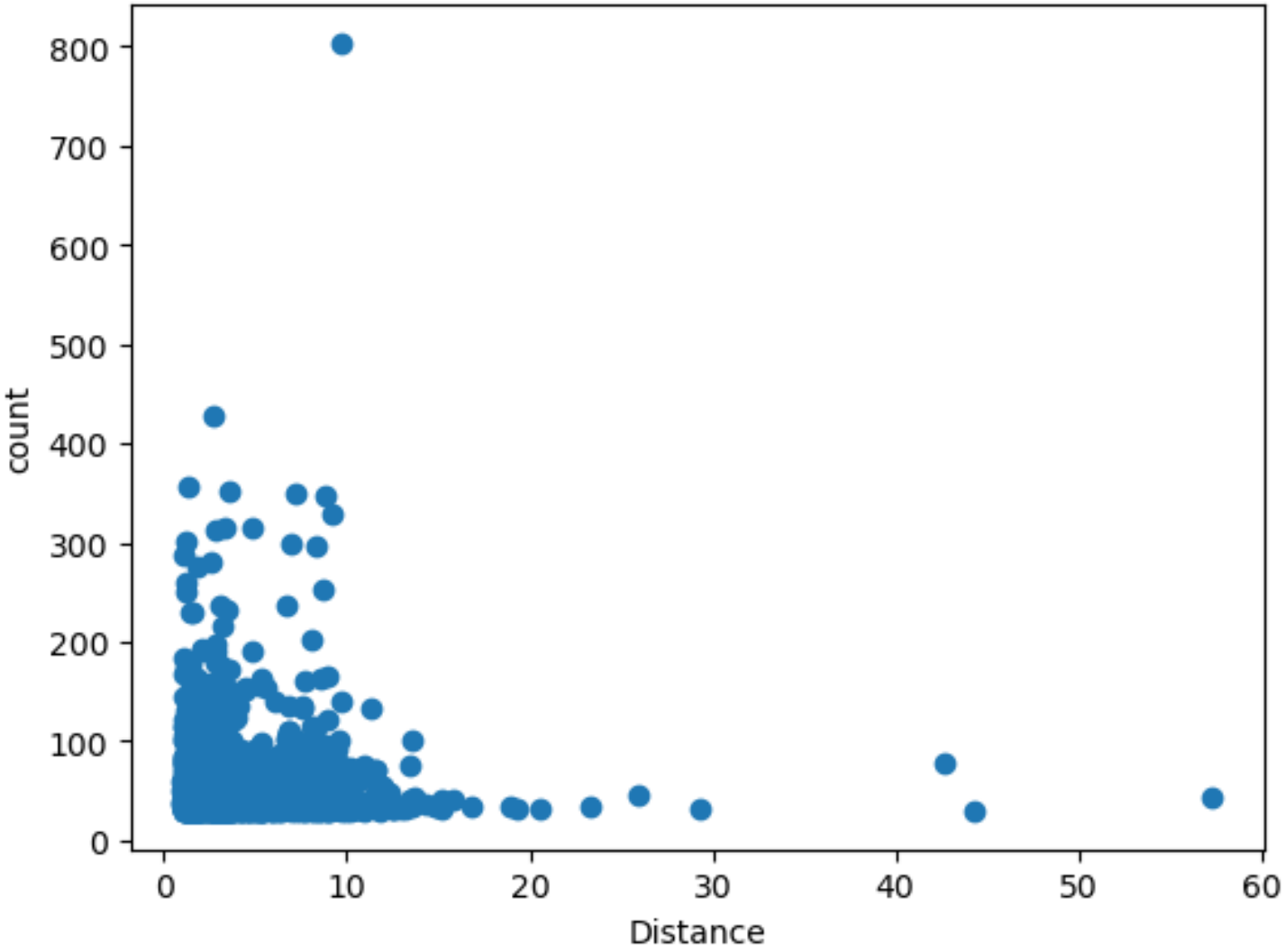
Répartition d'une simulation de jeux de données comportant des « hot spot » d'adresses administratives

Identification des biais géographiques

Méthode :

1. Filtre sur les adresses pertinentes

	Occurrence
Adresses différentes	788 765
1. Après filtre	2404



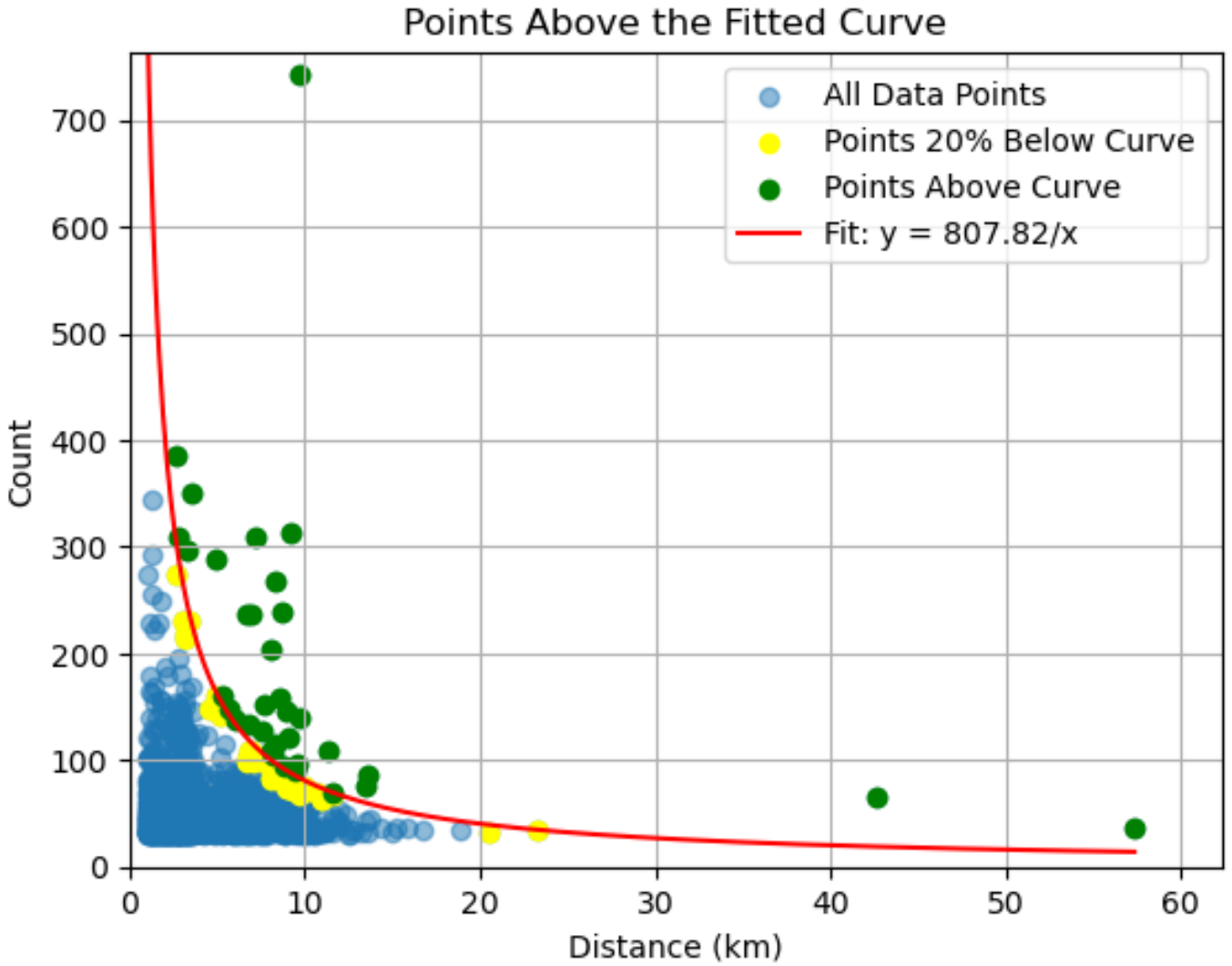
Occurrence des adresses selon leur distance

Identification des biais géographiques

Méthode :

2. Régression statistique

	Occurrence
Adresses différentes	788 765
1. Après filtre	2404
2. Après régression (80% de la valeur attendue)	64



Occurrence des adresses selon leur distance

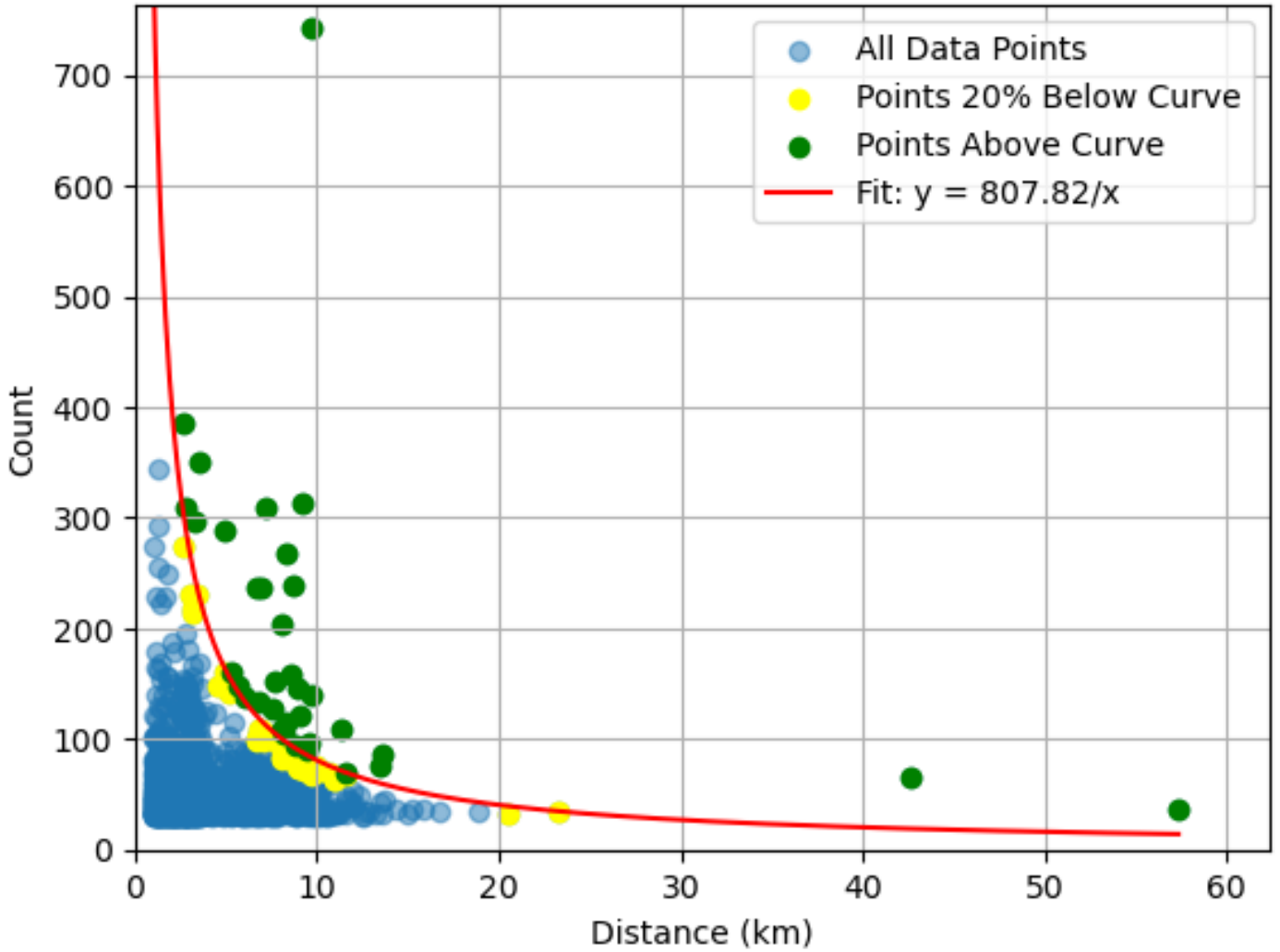
Identification des biais géographiques

Sur les 64 adresses identifiées,
45 sont des biais géographiques

soit **0.42** % de l'ensemble
des adresses postales correctes

Perspectives

Amélioration processus pour
détecter **faux positifs**



Occurrence des adresses selon leur distance

Enjeux et défis des données géospatiales pour l'oncologie

Conclusion

Des données géographiques à l'hôpital

La qualité de la donnée géographique est associées aux questions de parcours thérapeutique du patient – dépistage, soin, retour au travail etc. Mais aussi car les facteurs socioéconomiques et d'expositions ont un impact sur le cancer.

Des biais dans les données

L'identification des biais dans les adresses permet d'identifier les patients en grande précarité sociale, moins bien dépisté, afin d'adapter des plans de traitements.

Consolider la captation d'adresses

Qualité des données géographiques à travers un suivi spatio-temporelle des patients

– Nécessité des partenariat (CPAM , CHU, …)

ENJEUX ET DÉFIS DES DONNÉES

GÉOSPATIALES POUR L'ONCOLOGIE

GROUPE GEOCANCER

Cancéropôle Île-de-France

RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DU CANCER

Merci pour votre écoute

Joséphine Bocquet – Ingénieure Géomaticienne